M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems

Year: 2022 | Citations: 246 | Authors: Zeyu Cui, Jianxin Ma, Chang Zhou

---

Abstract

Industrial recommender systems have been growing increasingly complex, may involve \emph{diverse domains} such as e-commerce products and user-generated contents, and can comprise \emph{a myriad of tasks} such as retrieval, ranking, explanation generation, and even AI-assisted content production. The mainstream approach so far is to develop individual algorithms for each domain and each task. In this paper, we explore the possibility of developing a unified foundation model to support \emph{open-ended domains and tasks} in an industrial recommender system, which may reduce the demand on downstream settings' data and can minimize the carbon footprint by avoiding training a separate model from scratch for every task. Deriving a unified foundation is challenging due to (i) the potentially unlimited set of downstream domains and tasks, and (ii) the real-world systems' emphasis on computational efficiency. We thus build our foundation upon M6, an existing large-scale industrial pretrained language model similar to GPT-3 and T5, and leverage M6's pretrained ability for sample-efficient downstream adaptation, by representing user behavior data as plain texts and converting the tasks to either language understanding or generation. To deal with a tight hardware budget, we propose an improved version of prompt tuning that outperforms fine-tuning with negligible 1\% task-specific parameters, and employ techniques such as late interaction, early exiting, parameter sharing, and pruning to further reduce the inference time and the model size. We demonstrate the foundation model's versatility on a wide range of tasks such as retrieval, ranking, zero-shot recommendation, explanation generation, personalized content creation, and conversational recommendation, and manage to deploy it on both cloud servers and mobile devices.